# Spatially Stratified Sampling using Auxiliary Information for Geostatistical Mapping

M. G. Falk, R. J. Denham and K. L. Mengersen

**Abstract**

This paper presents a method of spatial sampling based on stratification by Local Morans I i calculated using auxiliary information. The sampling technique is compared to other design-based approaches including simple random sampling, systematic sampling on a regular grid, conditional Latin Hypercube sampling and stratified sampling based on auxiliary information, and is illustrated using two different spatial data sets. Each of the samples for the two data sets is interpolated using regression kriging to form a geostatistical map for their respective areas. The proposed technique is shown to be competitive in reproducing specific areas of interest with high accuracy.

## 1 Introduction

An objective of spatial sampling is the selection of sites so that geostatistical mapping can be performed to predict unsampled areas. The closer this map is to the true value, the better the sampling technique. Many different techniques exist for conducting design-based sampling. Reviews of such techniques are found in most statistical texts; in relation to spatial sampling, a seminal reference is Cressie (1993).

The motivation for this paper is work by Falk et al. (2010) who generate a map of uncertainty for deterministically estimated soil loss. The authors use a Bayesian melding technique (Poole and Raftery 2000) to estimate uncertainty in the Revised Universal Soil Loss Equation (RUSLE) (Renard et al. 1997). The RUSLE estimates hillslope erosion as a function of a set of environmental variables comprising rainfall, soil type, slope length and steepness, ground cover and supporting practice factors on a pixel by pixel basis. The authors quantify uncertainty in the form of prior distributions on the input and output (soil loss) factors and conclude that the slope steepness factor is the main contributor to total uncertainty. However, they note that the technique is too expensive for whole of catchment soil loss uncertainty estimation. An alternative to estimation of uncertainty for each pixel is to compute the uncertainties for a representative sample of pixels across the region of interest and interpolate

between these pixels. For this reason sampling techniques are required to generate a representative sample that, when interpolated, produces a reasonable soil loss uncertainty map. More specifically, because we are considering an uncertainty map, areas with high levels are those that we aim to reproduce with the highest accuracy.

This problem is such that we have auxiliary information about environmental variables for each pixel over the entire space we wish to sample, and this information is known to contribute in some way to the value of the outcome variable of interest. The popular methods of obtaining a sample of pixels include simple random sampling (SRS), systematic sampling on a regular grid (SYS), more informed sampling involving partitioning of the space and some innovative work using conditional Latin Hypercube sampling (Minasny and McBratney 2006), with Dobbie et al. (2008) providing a thorough review of spatial sampling approaches. While SRS and SYS ignore the auxiliary information, the latter two approaches use this information to stratify the space into more homogeneous regions that can be used to develop a more efficient sampling plan. SYS is commonly known to be the most efficient sampling approach (Müller 2007, for example) yet it often will miss important features (Sarsby 2000, for example). None of these approaches explicitly take advantage of the spatial nature of the auxiliary data. We suggest that incorporating this information will allow us to generate a sample that includes important features allowing accurate geostatistical mapping.

We consider here a version of stratified sampling based on local spatial autocorrelation. If we have auxiliary information for the whole area, and prior knowledge regarding the driving force behind the true parameter values, then it is worthwhile to incorporate any spatial relationships in the auxiliary data when designing a sampling technique. A sample is needed that sufficiently captures these characteristics of a map. We investigate whether sampling based on stratification by a local measure of spatial autocorrelation successfully achieves this goal.

Two case studies are used to illustrate the sampling techniques described in this paper. The first is the soil mapping data used in Minasny and McBratney (2006) from an area in the Hunter Valley of New South Wales, Australia. The second, the motivating data set for this paper, is from an uncertainty map of an area near Emerald in Queensland, Australia.

This paper is structured such that Section 2 firstly contains descriptions of the data sets, then outlines the sampling methods used, with particular attention paid to the method of spatial stratification. Section 2 concludes by explaining the interpolation of the samples over the entire surface and the measures to compare the sampling techniques. Section 3 gives the results of the sampling and subsequent geostatistical mapping for the case study areas. Section 4 provides a brief discussion of the results, and concludes by noting possible uses of the approach and areas for further improvement of the proposed technique.

# 2 Methods

## 2.1 Case Studies

The first study area is near Pokolbin in the Hunter Valley of New South Wales, Australia. The area considered is about about 11 $km^2$, with each pixel capturing 25 $m^2$ on the ground. The data set excludes pixels with land cover of water, roads and buildings (as it is to be used for digital soil mapping) which leaves 15 021 pixels in the data set (Minasny and McBratney 2006). We take two variables from the data: the Compound Topographic Index (CTI) and slope. The CTI values are calculated as a ratio between catchment area and slope, and are considered to be a measure of topographic moisture accumulation. Slope is slope angle, measured in degrees. Values of both variables are obtained from a Digital Elevation Model (DEM) of the area. The images are given in Figure 1. The white pixels are those excluded from the data set due to having a non-soil land cover listed above. A feature of this image is that the correlation between CTI and slope is negative, which distinguishes it from our second data set. The aim here is to reproduce the CTI map using auxiliary slope information.

The second case study is an uncertainty map for an area of approximately 14 $km^2$ near Emerald, Queensland, Australia, generated by Falk et al. (2010) using Bayesian melding on each individual 25 $m^2$ pixel in the image. The map is shown in the right-hand panel of Figure 2 and contains 21 888 pixels. The aim is to reproduce this uncertainty map using auxiliary information on the slope steepness factor (shown in the left-hand panel of Figure 2). By comparing the two images in Figure 2 we can see that areas of high uncertainty are found where the slope steepness is high. Thus the correlation between the slope steepness factor and uncertainty is positive, in contrast to the correlation structure found in the first data set. We use this prior information to design the spatially stratified sampling technique.

## 2.2 Sampling Approaches

As discussed in Section 1, a number of methods exist for selecting a sample of size $n$ from a population of $N$ measures of a variable, $Z$ say, of interest. In the context of this paper, $N$ is the total number of pixels in an image. Each method has different theoretical and practical properties, and to differing extents makes use of auxiliary information, $X$ say, about the spatial nature, magnitude and variability of $Z$.

Several different sampling techniques are considered and compared here to determine which method, when interpolated, reproduces an image most similar to the original. This is referred to as accuracy for the remainder of this article. More specifically, the image that is the most accurate for high values of the outcome variable is considered to be the better technique; this is in line with the original purpose behind this study, which was to develop a technique that accurately identified areas of large uncertainty.

We consider design-based sampling coupled with model-based geostatistical

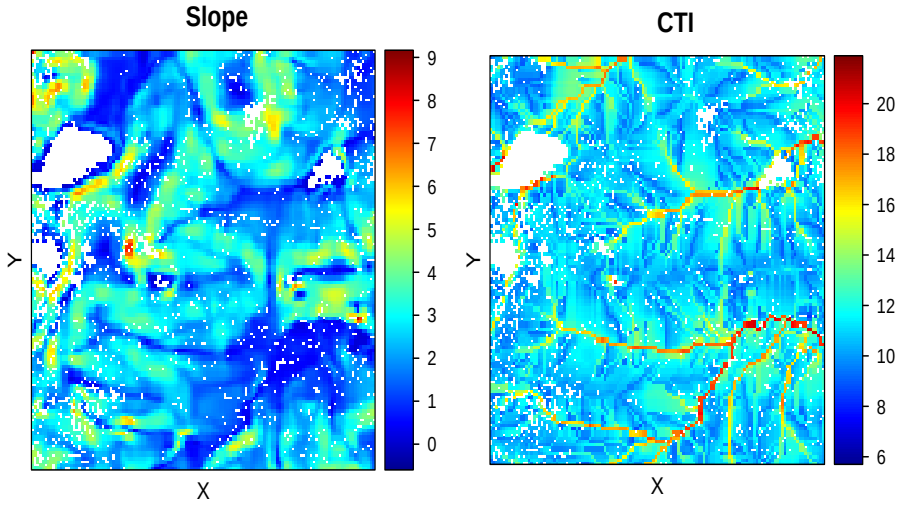


Figure 1: Left-hand panel: The slope in degrees of an area of approximately 11 $km^2$ near Pokolbin in the Hunter Valley, New South Wales, Australia. Right-hand panel: The map of Compound Topographic Index (CTI) values for the corresponding area. The white pixels are those excluded from the data set due to a land cover of water, roads or buildings as the original purpose of the data was for digital soil mapping (Minasny and McBratney 2006).

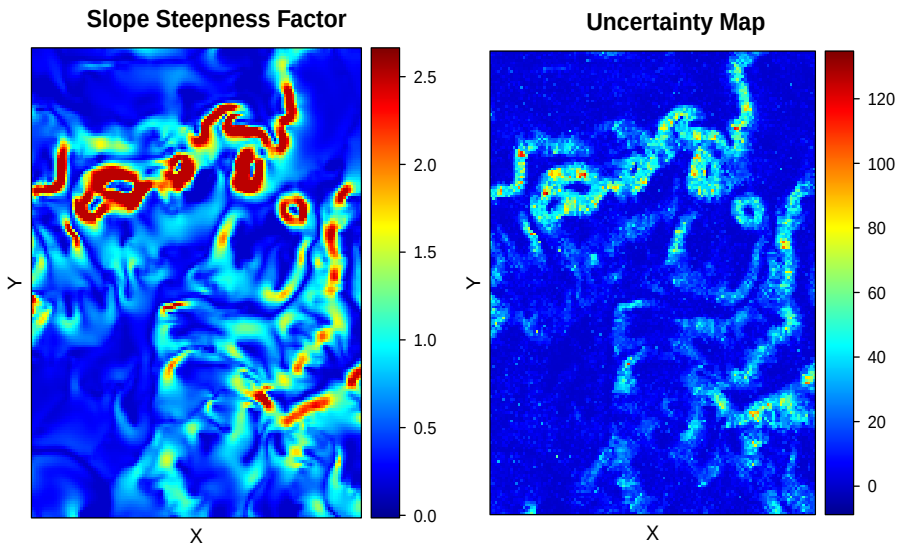


Figure 2: Left-hand panel: The slope steepness factor for an area of approximately 14 $km^2$ near Emerald, Queensland, Australia. Right-hand panel: Uncertainty map for the corresponding area (Falk et al. 2010), with higher values considered more uncertain.

mapping. A survey of the model-based sampling approach is found in Müller (2007). De Gruijter and Ter Braak (1992) provide a thorough comparison of design-based versus model-based sampling approaches, specifying advantages and disadvantages of each approach. Design-based sampling is unbiased and valid whether or not there is spatial autocorrelation (Brus and De Gruijter 1997), whereas model-based is more appropriate when prediction and spatial variation are of interest (Minasny and McBratney 2006). The aim of this paper is not to contribute to the discussion on which approach is better, but to propose a new design-based approach, and use model-based analysis for geostatistical mapping to compare it to commonly used sampling approaches.

The five sampling techniques considered include Simple Random Sampling (SRS), Systematic Sampling on a regular rectangular grid (SYS), Latin Hypercube Sampling conditioning on auxiliary information (cLHS), Stratified Sampling based on auxiliary information (StRS) and Stratified Sampling based on local spatial autocorrelation, using local Moran's $I$ (SpRS). We choose not to consider other stratified sampling techniques such as $k$-means clustering because cLHS gives better spatial coverage than this method (Minasny and McBratney 2006).

Each of the five methods will now briefly be outlined.

### 2.2.1 Simple Random Sampling

Simple random sampling (SRS) is the most common probabilistic sampling technique. From $N$ population values, we take $n$ random samples ($n < N$) of the measure of interest, $Z$. Sampling can be with or without replacement, but in the application of interest here, it typically without replacement because $n \ll N$. Choosing the sample size $n$ is an issue resolved by considering, among other issues, the required precision (Cochran 1963).

### 2.2.2 Systematic Sampling over a Regular Grid

Systematic Sampling can be over different types of grids. We consider a rectangular grid (SYS), however square, triangular and hexagonal grids are also used. SYS involves dividing a space into equal rectangular grids such that $n$ samples can then be taken from the centre of the grid, a random point within the grid, or from the intersection of grid lines, as in this paper. We note that while a triangular grid is considered more efficient, the rectangular grid is more convenient (Müller 2007), and for this reason it is the approach considered.

### 2.2.3 Conditional Latin Hypercube Sampling

Conditional Latin Hypercube sampling (cLHS) of an area is an approach that utilises prior knowledge in the form of complete auxiliary information. Given $N$ pixels with additional information regarding some variables $X$ over the entire space, the technique addresses the problem of selecting a sub-sample of size $n$ ($n \ll N$) such that it forms a Latin Hypercube, that is, the multivariate distribution of the sample is maximally stratified; see Minasny and McBratney

(2006) for details. Simply applying Latin Hypercube Sampling (LHS) to the auxiliary variables $X$ may result in a sample that is a combination of multivariate variables not present in the real world. Thus a search procedure is proposed by Minasny and McBratney (2006) that selects sites which form a Latin Hypercube in the observable space. In our examples, we use a single auxiliary variable, thus the sample is a one-dimensional Latin Hypercube sample of the auxiliary variable. The code provided with the original paper was used to produce the sample considered here.

#### 2.2.4 Stratified Sampling based on Auxiliary Information

Under Stratified Sampling (StRS), the population is divided into $k$ strata of size $N_k$, such that $N = \sum_k N_k$, based on the values of $X$. Then a simple random sample of $n_k$ is taken from each strata, such that $n = \sum_k n_k$. The sample sizes $n_k$ depend on the aim of the subsequent analysis. Common alternatives are equal sample sizes, values of $n_k$ proportional to stratum size and larger values of $n_k$ for strata of particular interest (Cochran 1963).

In our case study, we adopt the first approach of equal sample sizes $n_k$ but the stratum sizes $N_k$ may be considerably different. We know from previous research that auxiliary information, in particular slope steepness, over the entire map strongly influences the magnitude of the parameter value. Since the aim is to ensure that we capture the high areas of uncertainty for the Emerald data set (as these are found when the slope steepness factor is large), we stratify the population based on high and low slope steepness factors and take an equal sample from each. Approximately 5% of the pixels have a slope steepness factor greater than 2.2. In much the same way, we aim to capture the areas of large CTI values in the Pokolbin data, and these occur when slope is low. The sample is thus stratified based on slope values greater or less than 2.2, with approximately 45% of pixels in the latter stratum. Even though the value of 2.2 represents different measures (slope steepness factor units for the Emerald data set and degrees for the Pokolbin data set), the common value was chosen to illustrate both disparate and relatively similar stratum sizes. Note also that in both data sets considered, binary stratification is used. However, the method generalises to multiple strata in an obvious way.

#### 2.2.5 Stratified Sampling based on Local Spatial Autocorrelation of Auxiliary Information

Here we consider a sampling scheme that builds on SRS and StRS by adding a spatial component. We denote this spatially stratified sampling scheme by SpRS. A global measure of spatial autocorrelation proposed by Moran (1950) is obtained for the auxiliary information, here slope steepness. This is defined as

$$I = \frac{N \sum_i \sum_j w_{ij}(x_i - \bar{x})(x_j - \bar{x})}{S_0 \sum_i (x_i - \bar{x})^2},$$

where $N$ is the total number of pixels in an image indexed by $i$ and $j$, $\overline{x}$ is the mean of the auxiliary variable of interest, $S_0 = \sum_i \sum_j w_{ij}$ and $w_{ij}$ is a matrix of spatial weights. The weight $w_{ij}$ is a binary matrix with a value of one in position $(i, j)$ whenever the observation $i$ is in the neighbourhood of observation $j$. Without loss of generality, we adopt a rook design in which a neighbourhood comprises pixels with a common boundary.

The global measure of autocorrelation assumes that the image is homogeneous. If this may not be the case then Anselin (1995) suggested a local measure of spatial autocorrelation, Local Moran's $I_i$, to identify areas of non-stationarity or 'hot spots'. For the $i$th pixel, this measure is defined as

$$I_i = \frac{(x_i - \bar{x})}{\sigma^2} \sum_{j=1, j\neq i}^{N} w_{ij}(x_j - \bar{x}),$$

where $\sigma^2$ is the variance of $x$. Anselin (1995) also showed that the sum of all Local Moran's $I_i$ values is equal to the global Moran statistic.

Thus a stratified sample can be generated based on the $I_i$ values. Following earlier arguments in this paper we stratify the image in a way that captures an equal amount of high and low $I_i$ values. This approach ensures that we identify those areas with high and low autocorrelation, that is, areas that are quite similar and different, in order to reproduce the key features of the map.

Based on expert opinion, the threshold for the strata was set at a Local Moran's $I_i$ of 2; values above this indicate very strong significant spatial autocorrelation (p ¡ 0.0001) with approximately 15% and 10% of pixels in this stratum for the Pokloblin and Emerald data sets respectively. Two strata were chosen for consistency with StRS.

## 2.3 Geostatistical Mapping Of The Surface

Kriging is a common approach to predicting a random variable, $Z$, at unsampled location $s_0$, that is $\hat{z}(s_0)$, given a sample $z(s_1), z(s_2), ..., z(s_n)$ (Webster and Oliver 2007). For this paper we consider regression kriging which allows the use of auxiliary variables, $X$, known over the entire space; and the relationship between auxiliary variables and the primary variable, $Z$, for interpolation. Using similar notation to Hengl et al. (2007), regression kriging (also known as Kriging with External Drift or Universal Kriging) predicts $z(s_0)$ at the same resolution as the sample by

$$\begin{aligned} \hat{z}(s_0) &= \hat{m}(s_0) + \hat{e}(s_0) \\ &= \sum_{k=0}^{p} \hat{\beta}_k \cdot q_k(s_0) + \sum_{i=1}^{n} w_i(s_0) \cdot e(s_i); \qquad q_0(s_0) = 1, \end{aligned}$$

where $\hat{\beta}_k$ are the least squares estimates of the regression model coefficients, $p$ is the number of predictors, $q_k(s_0)$ are the auxiliary variable values at the target locations, $w_i(s_0)$ are kriging weights determined by the spatial covariance

function and $e(s_i)$ are the regression residuals. This approach uses regression to fit the response to explanatory variable(s) and then residuals are fitted using simple kriging with an expected value of 0. Regression kriging is the combination of these two techniques. Although we have indicated anisotropy by identifying strata, kriging is conducted over all $Z$ values rather than within strata for consistency with all sampling approaches considered. The reader is referred to Hengl et al. (2004) for additional information on regression kriging.

The kriging was carried out in R (R Development Core Team 2008), using the *gstat* package, with spherical variogram models as these fitted the sampled data adequately and are commonly used theoretical variogram models. Simple interpolation, other forms of kriging such as cokriging, or Markov Random fields could have also been used to generate the surface (Marin and Robert 2007); however regression kriging was designed for the situation described (Knotters et al. 1995), was easy to implement and sufficiently incorporates the spatial relationships in the data sets considered here.

### 2.4 Measures for Comparing Sampling Methods

Two measures are used for comparing the sampling methods presented in this paper. The first is the root mean square error (RMSE). The mean squared error (MSE) is the squared difference between true values and estimates averaged over all samples. The RMSE is the square root of the MSE. The RMSE quoted in Section 3 are averaged over the multiple runs of each procedure to quantify the difference of predictions from the true values. The second measure is the average standard error (ASE). The ASE is the standard error of the predictions, averaged over all the pixels. The ASE listed in Section 3 are averaged over the multiple runs for each sampling method. If the ASE is close to the RMSE, then the method is asserted to be accurately capturing the variability in the predicted surface. If the ASE is greater than the RMSE, the variability is being overestimated whereas if the ASE is less than RMSE, the variability is being underestimated.

## 3 Results

For both the Pokolbin and Emerald data sets a sample of size $n = 1000$, from total pixel sizes of $N_P = 15\,021$ and $N_E = 21\,888$ respectively, was chosen based on each of the five techniques listed above. The sample size was chosen somewhat arbitrarily, however acknowledging the sample variance (approximately 200) and chosen margin of error (0.8 say, implying we can reasonably accept a 0.8 difference from the true value) for the motivating Emerald data set, the theory of Cochran (1963) supports this choice. The same sample size was used for the Pokolbin data set for consistency, although the sample variance is considerably smaller (approximately 3). Sample sizes for the systematic samples on regular grids were slightly greater than 1000 due to the restrictions placed on grid locations given the data is in the form of pixels from an image. The

results are now compared with respect to the accuracy of reproduction of the original images from the two study areas.

## 3.1 Pokolbin

The upper left panel of Figure 3 shows the map of true CTI values for the Pokolbin area. The kriged results from each sample are given in the other panels of Figure 3. It can be seen that each of these identify some of the patterns present in the original true image, however some of the approaches capture features missed by others. For example, in the lower right corner of the original image (labelled True in Figure 3), the high areas are being captured better by the SpRS as compared to the other samples.

The residual plots from the sampling techniques are given in Figure 4. For total CTI values, all the samples are very similar, with possibly the stratified samples performing with slightly lower accuracy. CTI values are then split into low (CTI $<$ 10), medium (10 $\leq$ CTI $\leq$ 15) and high (CTI $>$ 15) categories. When we consider high CTI values, perhaps SpRS provides a slightly more accurate image.

The samples for the techniques considered above may not reflect the 'average' nature of samples generated from that technique. Thus multiple runs for each of the sampling techniques (except for SYS, which has one design) were conducted with the RMSE and ASE noted. The results are given in Table 1. This gives the SYS as the best technique in terms of lowest RMSE for reproduction of overall values. However, the aim was to reproduce high parameter values and the SpRS does so with the most accuracy; see Table 1. Additionally, the increased overall RMSE for SpRS is not much higher than the other approaches considered. The average standard errors (ASE) are also given in Table 1. As a consequence of the stratification and reproduction of high CTI values, the SpRS gives the largest difference between RMSE and ASE for all CTI levels. The ASE values will be discussed further when considering the Emerald data.

## 3.2 Emerald

The uncertainty map that we are aiming to reproduce is given in the upper left panel of Figure 5. The uncertainty maps generated by kriging the samples obtained using the five approaches, are given in the remaining panels of Figure 5.

As expected the kriged maps demonstrate considerable smoothing compared to the original. All five kriged maps present very similar patterns to that in the original map, however the stratified samples appear to be capturing the higher values within the boxed areas of the original image (upper left panel, Figure 5). Figure 6 contains the residuals of the four sampling techniques for the uncertainty overall, and then broken up into low (uncertainty $<$ 2), medium (2 $\leq$ uncertainty $\leq$ 15) and high (uncertainty $>$ 15). No single sampling method stands out visually as better than the others in terms of overall uncertainty. Perhaps the SpRS is performing very slightly better for high uncertainty, but

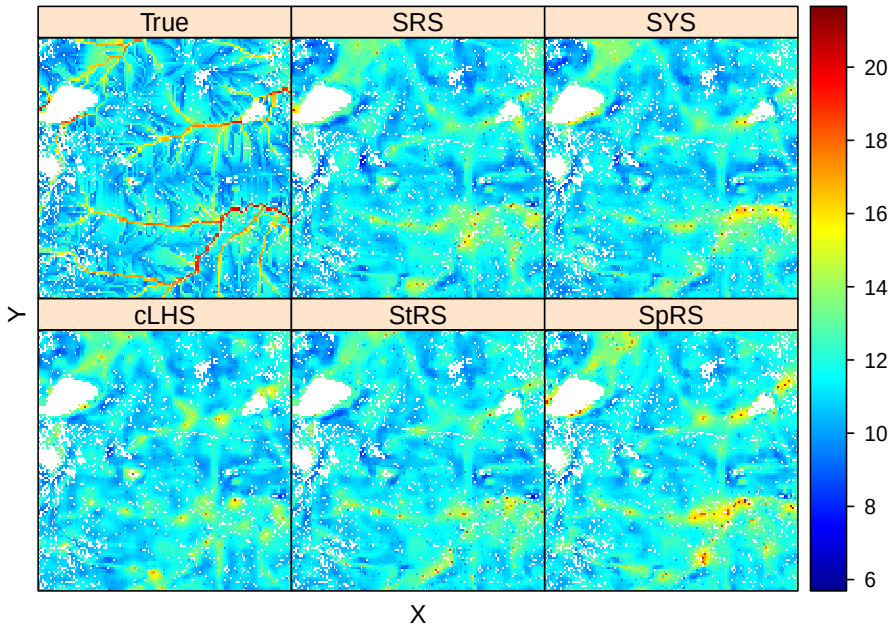


Figure 3: Comparison of sampling techniques for the Pokolbin data, which were then kriged to reproduce the image of CTI values. Upper left panel: The true map of CTI values. Upper middle panel: simple random sample. Upper right panel: systematic sample using a regular rectangular grid. Lower left panel: conditional Latin Hypercube sample. Lower middle panel: stratified by the auxiliary slope information. Lower right panel: stratified by Local Moran's $I_i$ of auxiliary slope information. The high CTI values are being captured by the SpRS.

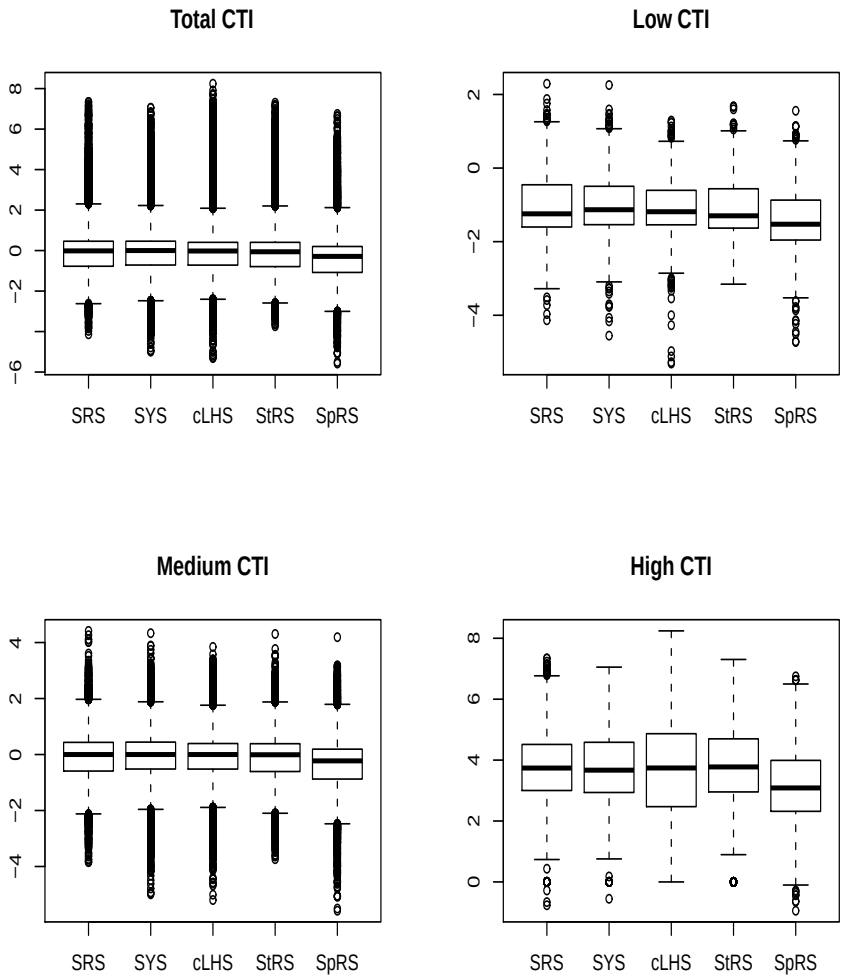


Figure 4: Residuals under the five sampling techniques for the Pokolbin data. Upper left panel: residuals for total CTI. Upper Right panel: residuals for low ($< 10$) CTI. Lower left panel: residuals for medium ($\geq 10$ and $\leq 15$) CTI. Lower right panel: residuals for high ($> 15$) CTI.

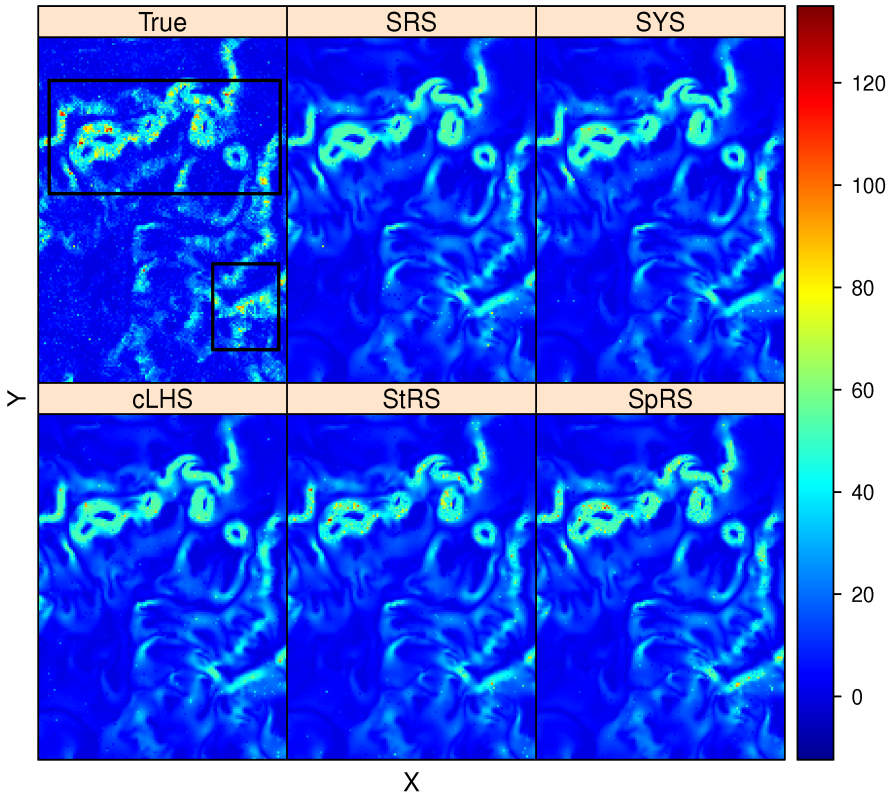


Figure 5: Comparison of sampling techniques, which were then kriged to reproduce the soil loss uncertainty image for the Emerald data set. Upper left panel: The true map of uncertainty values. Upper middle panel: simple random sample. Upper right panel: systematic sample using a regular rectangular grid. Lower left panel: conditional Latin Hypercube sample. Lower middle panel: stratified by the auxiliary slope steepness factor information. Lower right panel: stratified by Local Moran's $I_i$ of auxiliary slope steepness factor information. The stratified samples are capturing the higher values inside the boxed areas of the true image.

poorly for low and medium uncertainty, as compared to the other sampling techniques.

As with the Pokloblin data set, multiple runs of the sampling procedures (except for SYS, which has one design) were conducted to ensure 'average' samples were presented for the Emerald data set. The results from the repeated sampling are summarised in Table 2. Based on the RMSE values for the kriged surfaces, overall, the best sampling techniques were StRS, SpRS, and SYS, followed by cLHS, with SRS having the lowest accuracy; see Table 2.

The aim in this case study was not only to reproduce the map, but specifically identify accurately those areas with high uncertainty. Considering the RMSE estimates for low, medium and high uncertainty, as given in Table 2, the SpRS gives the lowest RMSE for high uncertainty. This is at the detriment of the low and medium uncertainty levels. The performance of SpRS for high levels is perhaps not surprising in this example, given the relatively intensive sampling of the corresponding stratum. The StRS, SYS, cLHS and SRS outperform the SpRS for low and medium uncertainty levels.

The average standard errors (ASE) are similar to the RMSE for all techniques except the stratified samples for both Emerald and Pokolbin data, however the difference is more noticeable in the Emerald data set. The SpRS overestimates the variability as the ASE is greater than the RMSE for all uncertainty levels in the Emerald data. Interestingly, the SpRS underestimates the variability for some levels in the Pokolbin data. However, for high levels of both the Pokolbin and Emerald data, RMSE and ASE are the closest for SpRS. This indicates that for high levels SpRS not only has the lowest RMSE but also estimates the variability most accurately.

# 4 Discussion

This paper has compared several different techniques for sampling an area to reproduce an original image using kriging. The approaches considered were Simple Random Sampling (SRS), systematic sampling on regular rectangular grid (SYS), Latin Hypercube Sampling conditioning on auxiliary information (cLHS), Simple Stratified Sampling based on auxiliary information (StRS) and Spatial Stratified Sampling based on spatial autocorrelation of auxiliary information (SpRS). The SpRS technique proposed here stratifies an image based on Local Moran's $I_i$ of auxiliary information known to be affecting the parameter of interest. This approach attempts to select sites in strata that are both similar and different based on auxiliary variable spatial autocorrelation. SpRS was shown to accurately capture high parameter levels in a map, for the two given case studies, better than the other sampling techniques considered. The comparisons are reported to the first decimal place which represents an acceptable level of accuracy.

We see the benefit of such a sampling technique when remotely sensed data can be used as the auxiliary variable information. This is because such data is usually available across the entire space, allowing for calculation of local spatial

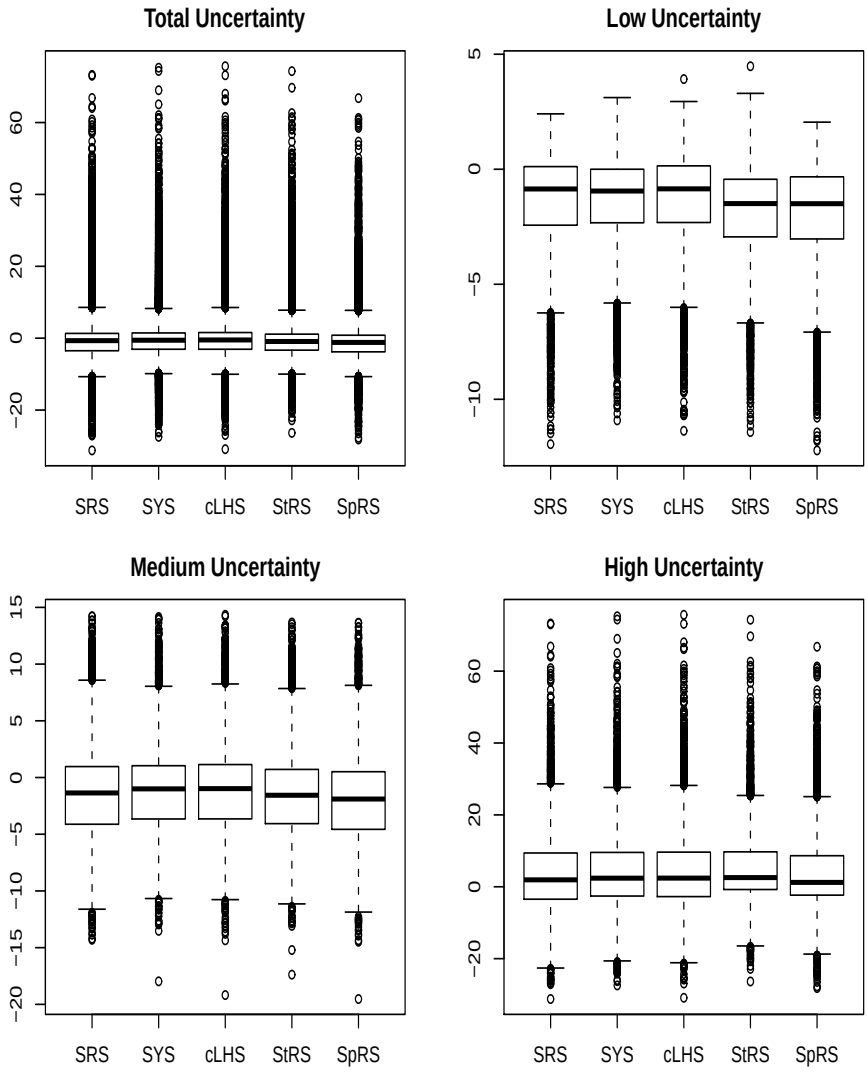


Figure 6: Residuals under the five sampling techniques for the Emerald image of uncertainty estimates. Upper left panel: residuals for total uncertainty. Upper Right panel: residuals for low ($< 2$ units) uncertainty. Lower left panel: residuals for medium ($\geq 2$ and $\leq 15$) uncertainty. Lower right panel: residuals for high ($> 15$) uncertainty.

Table 1: The mean of RMSE and average standard error (ASE) for 1000 multiple runs (except for SYS) of each sampling technique for the Pokolbin data set. Low CTI is $< 10$ units; medium CTI is $\geq 10$ and $\leq 15$; high CTI is $> 15$.

| Sampling Technique | CTI Level | RMSE | ASE |
|---|---|---|---|
| SRS | Overall | 1.3198 | 1.2365 |
| | Low | 1.3183 | 1.2488 |
| | Medium | 0.9193 | 1.2459 |
| | High | 3.9488 | 1.2494 |
| SYS | Overall | 1.3118 | 1.2696 |
| | Low | 1.3026 | 1.2601 |
| | Medium | 0.9262 | 1.2701 |
| | High | 3.8927 | 1.2853 |
| cLHS | Overall | 1.3160 | 1.2597 |
| | Low | 1.3186 | 1.2617 |
| | Medium | 0.9122 | 1.2592 |
| | High | 3.9514 | 1.2623 |
| StRS | Overall | 1.3140 | 1.3025 |
| | Low | 1.3244 | 1.3107 |
| | Medium | 0.9191 | 1.3018 |
| | High | 3.9079 | 1.2915 |
| SpRS | Overall | 1.3367 | 1.5331 |
| | Low | 1.5414 | 1.5173 |
| | Medium | 0.9821 | 1.5407 |
| | High | 3.6013 | 1.4559 |

Table 2: The mean of RMSE and average standard error (ASE) for 1000 multiple runs (except for SYS) of each sampling technique for the Emerald data set. Low uncertainty is $< 2$ units; medium uncertainty is $\geq 2$ and $\leq 15$; high uncertainty is $> 15$.

| Sampling Technique | Uncertainty Level | RMSE | ASE |
|---|---|---|---|
| SRS | Overall | 6.8080 | 6.5733 |
| | Low | 2.5473 | 6.5730 |
| | Medium | 4.0410 | 6.5718 |
| | High | 11.9151 | 6.5766 |
| SYS | Overall | 6.6970 | 6.8277 |
| | Low | 2.4277 | 6.8356 |
| | Medium | 3.7954 | 6.8037 |
| | High | 11.8518 | 6.8649 |
| cLHS | Overall | 6.8050 | 7.0942 |
| | Low | 2.4715 | 7.1371 |
| | Medium | 3.9668 | 7.0912 |
| | High | 11.9710 | 7.0571 |
| StRS | Overall | 6.5532 | 13.2259 |
| | Low | 2.6199 | 13.5416 |
| | Medium | 3.8749 | 13.5363 |
| | High | 11.4423 | 12.3320 |
| SpRS | Overall | 6.6385 | 12.6768 |
| | Low | 2.8398 | 12.9567 |
| | Medium | 4.1294 | 12.9526 |
| | High | 11.4131 | 11.8832 |

correlation and designing the subsequent sampling by stratification over those values. Additionally, if one wanted to select sample sites to accurately predict a map (particularly certain values) with restricted resources, this technique may be appropriate. Also a similar context may evolve similar to the motivation for this paper whereby we had a computationally expensive procedure. Then we could sample using SpRS to achieve the desired results. There is naturally some extra computational expense in the calculation of Local Moran's $I_i$.

Of course, under other circumstances the different approaches may give different relative performances. For example, a criticism of both StRS and SpRS is the setting of thresholds for the strata. In our case, strata were set to ensure parameter values of particular interest were sampled, but strata may also be determined to simplify data collection or due to known regions of homogeneity (Cochran 1963). We have not given a transparent method by which this can be set, but this is an area that can be considered for further research.

Finally, other sampling techniques could be considered such as adaptive sampling schemes (for example, Marchant and Lark 2006). For our objectives, which were designing an efficient, practical and accurate means of sampling for map reproduction, which makes use of available spatial and auxiliary information, stratification by Local Moran's $I_i$ was found to be adequate.

# References


Anselin, L. (1995). Local indicators of spatial association - LISA. *Geographical Analysis 27*(2), 93–115.

Brus, D. J. and J. J. De Gruijter (1997). Random sampling or geostatistical modelling? Choosing between design-based and model-based sampling strategies for soil (with Discussion). *Geoderma 80*(1-2), 1–44.

Cochran, W. G. (1963). *Sampling Techniques*. New York: Wiley.

Cressie, N. (1993). *Statistics for Spatial Data* (revised ed.). New York: John Wiley & Sons, Inc.

De Gruijter, J. J. and C. J. F. Ter Braak (1992). Design-based versus model-based sampling strategies: Comment on R. J. Barnes' "Bounding the required sample size for geologic site characterization". *Mathematical Geology 24*(7), 859–864.

Dobbie, M. J., B. L. Henderson, and D. L. Stevens, Jr. (2008). Sparse sampling: Spatial design for monitoring stream networks. *Statistics Surveys 2*, 113–153.

Falk, M. G., R. J. Denham, and K. L. Mengersen (2010). Estimating uncertainty in the Revised Universal Soil Loss Equation via Bayesian melding. *Journal of Agricultural, Biological and Environmental Statistics*. In Press.

Hengl, T., G. B. M. Heuvelink, and D. G. Rossiter (2007). About regression-kriging: From equations to case studies. *Computers and Geosciences 33*(10), 1301–1315.

Hengl, T., G. B. M. Heuvelink, and A. Stein (2004). A generic framework for spatial prediction of soil variables based on regression-kriging. *Geoderma 120*(1-2), 75–93.

Knotters, M., D. J. Brus, and J. H. O. Voshaar (1995). A comparison of kriging, co-kriging and kriging combined with regression for spatial interpolation of horizon depth with censored observations. *Geoderma 67*(3-4), 227–246.

Marchant, B. P. and R. M. Lark (2006). Adaptive sampling and reconnaissance surveys for geostatistical mapping of the soil. *European Journal of Soil Science 57*(6), 831–845.

Marin, J.-M. and C. P. Robert (2007). *Bayesian Core: A Practical Approach to Computational Bayesian Statistics*. New York: Springer.

Minasny, B. and A. B. McBratney (2006). A conditioned Latin hypercube method for sampling in the presence of ancillary information. *Computers and Geosciences 32*(9), 1378–1388.

Moran, P. A. P. (1950). Notes on continuous stochastic phenomena. *Biometrika 37*, 17–23.

Müller, W. G. (2007). *Collecting Spatial Data* (3rd ed.). Heidelberg: Springer Verlag.

Poole, D. and A. E. Raftery (2000). Inference for deterministic simulation models: The Bayesian melding approach. *Journal of the American Statistical Association 95*(452), 1244–1255.

R Development Core Team (2008). *R: A Language and Environment for Statistical Computing*. Vienna, Austria: R Foundation for Statistical Computing. ISBN 3-900051-07-0.

Renard, K. G., G. R. Foster, G. A. Weesies, D. K. McCool, and D. C. Yoder (1997). *Predicting soil erosion by water: A guide to conservation planning with the Revised Universal Soil Loss Equation (RUSLE)*. Washington, DC, USA: Agriculture Handbook. No. 703. U.S. Department of Agriculture.

Sarsby, R. W. (2000). *Environmental Geotechnics*. London: Thomas Telford Ltd.

Webster, R. and M. A. Oliver (2007). *Geostatistics for Environmental Scientists* (2nd ed.). Chichester, U.K.: John Wiley & Sons, Ltd.